Bias away from the null due to miscounted outcomes? A case study on the TORCH trial

Muff, Stefanie ; Puhan, Milo A ; Held, Leonhard

Abstract: Count outcomes occur in virtually all disciplines, such as medicine, epidemiology or biology, but they often contain error. Recently, it has been shown that self-reported numbers of exacerbations of Chronic Obstructive Pulmonary Disease patients can be considerably miscounted. Motivated by this result, we reanalysed data from the Towards a Revolution in Chronic Obstructive Pulmonary Disease Health trial, a large randomized controlled trial with the self-reported number of exacerbations of Chronic Obstructive Pulmonary Disease patients as outcome. To adjust for miscounting error in the response of Poisson and (zero-inflated) negative binomial models, we introduce novel, general methodology. The key idea is to formulate a zero-inflated negative binomial model to capture the error mechanism. This parametric approach automatically circumvents drawbacks of previously suggested methodology that treats miscounted outcomes in the misclassification framework. Prior information for the response error model parameters was elicited from validation data of an external study and adaptively weighted to account for potential prior-data conflict. The results of the Bayesian hierarchical modelling approach indicated that the treatment effect has been overestimated in the original study. However, closer inspection revealed that this unexpected result was an artefact of an unaccounted time dependency of the treatment effect.

DOI: https://doi.org/10.1177/0962280217694403

Bias away from the Null due to miscounted outcomes? A case study on the TORCH trial

Statistical Methods in Medical Research
XX(X):4–42

DOI: 10.1177/ToBeAssigned
www.sagepub.com/

Stefanie Muff[1,2], Milo A. Puhan[1] and Leonhard Held[1]

Prepared using sagej.cls

Abstract

Count outcomes occur in virtually all disciplines, such as medicine, epidemiology or biology, but they often contain error. Recently it has been shown that self-reported numbers of exacerbations of Chronic Obstructive Pulmonary Disease (COPD) patients can be considerably miscounted. Motivated by this result, we reanalyzed data from the Towards a Revolution in COPD Health (TORCH) trial, a large randomized controlled trial with the self-reported number of exacerbations of COPD patients as outcome. To adjust for miscounting error in the response of Poisson and (zero-inflated) negative binomial models we introduce novel, general methodology. The key idea is to formulate a zero-inflated negative binomial model to capture the error mechanism. This parametric approach automatically circumvents drawbacks of previously suggested methodology that treats miscounted outcomes in the misclassification framework. Prior information for the response error model parameters was elicited from validation data of an external study, and adaptively weighted to account for potential prior-data conflict. The results of the Bayesian hierarchical modelling approach indicated that the treatment effect has been overestimated in the original study. However, closer inspection revealed that this unexpected result was an artefact of an unaccounted time-dependency of the treatment effect.

Keywords

Miscounting error; response error; count outcome; zero-inflated negative binomial regression; Bayesian analysis; randomized clinical trial; prior weighting;

1 Introduction

Investigating the effects of measurement error (ME) on the parameter estimates of regression models has a long tradition in the statistical literature[1–6]. Bias induced by ME can be classified into attenuation (bias towards zero) and reverse attenuation (bias away from zero) effects. The vast majority of literature on ME in regression focusses on error in the covariates, which is also reflected by the attention given to it by recent monographs on error modelling[5–7].

In contrast to the covariates, which are not required to obey any distributional assumptions and are assumed to be error-free in standard regression methods, variability in the response is allowed and modelled via the likelihood of the regression model. In linear regression, for instance, unbiased, additive, homoscedastic ME in the response of a linear model is simply absorbed in the variance of the distribution and thus requires no additional modelling efforts[5;8]. For heteroscedastic error in a continuous outcome, weighted regression or generalized least squares methods can be used[9], and methods for biased continuous outcome have been proposed as well[10–12]. On the other hand, there is no variance term in logistic regression, for example, that absorbs misclassification error in the response, and this case thus needs specific treatment. Various methods have been proposed for this problem[13–16], among others an EM algorithm to

[1] Epidemiology, Biostatistics and Prevention Institute (EBPI), University of Zurich, Hirschengraben 84, 8001 Zurich, Switzerland
[2] Department of Evolutionary Biology and Environmental Studies (IEU), University of Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland

Corresponding author:
Stefanie Muff, Epidemiology, Biostatistics and Prevention Institute (EBPI), University of Zurich, Hirschengraben 84, 8001 Zurich, Switzerland
Email: stefanie.muff@uzh.ch

Prepared using sagej.cls

recover unbiased estimates of the odds ratios and their variances[17] and a Bayesian approach[18].

In clinical trials, exacerbation numbers are frequently used as a response variable, for example in the TORCH study[19], where the self-counted rate of moderate COPD exacerbations was included as a secondary endpoint. Recently, Siebeling, Frei and co-authors[20;21] have shown in the context of the International Collaborative Effort on Chronic Obstructive Lung Disease: Exacerbation Risk Index Cohorts (ICE COLD ERIC) study that self-counted exacerbation numbers of COPD patients may contain considerable *miscounting error*. We therefore focus in this paper on the effects of miscounted outcomes in standard regression models, namely Poisson, negative binomial (NB), or zero-inflated negative binomial (ZINB) models. So far miscounted outcomes have been treated within the misclassification framework[22]. Although this approach is extremely flexible, as all probability entries of the misclassification matrix can, in principle, be estimated separately from validation data, it has also a number of disadvantages, such as the large number of parameters that need to be estimated, or that some additional assumptions about these entries are required, *e. g.* decaying misclassification probabilities for more dissimilar values according to some functional form[23]. An additional problem is that the dimension of the misclassification matrix automatically constrains the range of possible true counts.

To overcome such difficulties, we propose a parametric miscounting error model and formulate a ZINB regression model for the distribution of observed counts as a function of the true counts. This error model is then embedded in a Bayesian hierarchical modelling framework, so that posterior marginals of the regression and error models can be estimated jointly. In the simplest setup the error model is independent of the covariates, in which case the error is *non-differential* and

the observed response is a *surrogate* for the true response[5;24]. However, it can be useful to formulate the response error model in more generality by allowing for covariate-dependency[15;16;18;22], thus *differential* error. Implications and effects of differential and non-differential response error and how to model it will be discussed.

In the application to the TORCH trial we used the information from the ICE COLD ERIC study to estimate prior distributions for the error model parameters, *i. e.* the latter provided us with *validation data*. However, even when validation data are collected under similar conditions as the data of interest, there is some potential for a prior-data conflict[25] when transporting error model parameters from one study to another. A recently suggested prior weighting approach[26] was therefore used to account for such problems. We used a non-differential model for the miscounting process, which led to a smaller estimated treatment effect with respect to the results from the original study, indicating that the effect has previously been overestimated. However, this result is in contrast to theoretical predictions. Closer inspection of the original regression model from the TORCH trial revealed that the estimated treatment effect weakened over the duration of medication. Extending the regression model by adding an interaction of 'time under treatment' with 'treatment' changed the direction of the error correction effect. In addition, the validation data allowed to estimate an error model with an explicit sex-dependency, which allows for a potentially different reporting behavior of males and females.

This paper is organized as follows. We will start by introducing the TORCH trial[19]. We will then describe the validation dataset extracted from the ICE COLD ERIC study and illustrate that a ZINB error model for the miscounted exacerbations numbers captures the error process well. The methods section

then generalizes ZINB error modelling, which contains Poisson and NB models as a special case. We discuss potential effects of error in the response, and describe how the error and regression models can be integrated into a Bayesian hierarchical model. We then apply the methodology in the subsequent section to our motivating example. The final section provides some conclusions, and we discuss the importance, but also possible difficulties, of error modelling.

2 Case study: the TORCH trial

The TORCH trial[19] was a large trial of pharmacotherapy in patients with COPD that lasted for 3 years. The study included 6112 patients in the efficacy population, of which $n_1 = 1524$ received a placebo and $n_2 = 1533$ a combination treatment (salmeterol plus fluticasone). Another 1521 patients received only salmeterol and 1534 received only fluticasone, but these treatment arms were not included in our analyses. The primary objective of the study was to demonstrate a significant reduction in all-cause mortality in COPD subjects that obtained the combination treatment, compared with the placebo group. The rate of moderate COPD exacerbations was included as a secondary endpoint, on which our interest centers in this example. The data from the TORCH study were provided and accessed through the SAS Solutions on Demand secure portal (`https://researchenvadmin.ondemand.sas.com`). All statistical analyses were carried out in the provided Clinical Trial Data Transparency Research Environment, from where results could then be exported.

The frequency of exacerbations were analyzed in Calverley *et al.*[19] using a NB model with treatment $x_i \in \{0, 1\}$ of patient i, adjusted for region of recruitment (Eastern Europe, Western Europe, USA, Asia and Pacific, Other), age, sex, baseline smoking status (yes/no), BMI, number of exacerbations in the 12 months

prior to screening (categorized as 0, 1, ≥ 2), and baseline disease severity. These confounder variables were summarized in the vector $\boldsymbol{z}_i$. To account for inter-individual differences in the time under treatment, the logarithm of the time $\log(t_i)$ during which patient i received the allocated treatment in the study was included as an offset variable[27], thus the original regression model was given as

$$y_i \sim \mathsf{NB}\left(\exp(\log(t_i) + \beta_0 + x_i\beta_x + \boldsymbol{z}_i\boldsymbol{\beta}_z), \theta\right) \ . \tag{1}$$

The probability mass function of the $\mathsf{NB}(\mu, \theta)$ distribution with overdispersion parameter θ, expected value $\mathsf{E}(y) = \mu$ and variance $\mathsf{Var}(y) = \mu(1 + \mu/\theta)$ is given in Appendix A. Note that the overdispersion parameter θ is inversely related to the variance $\mathsf{Var}(y)$, and in particular a Poisson distribution with $\mathsf{E}(y) = \mathsf{Var}(y)$ is obtained for $\theta \to \infty$.

The regression parameters β_0, β_x and $\boldsymbol{\beta}_z$ in (1) are the intercept, the treatment effect and the parameters of the remaining covariates, respectively. The rate ratio for treatment vs. placebo was estimated as $\exp(\hat{\beta}_x) = 0.75$ with a 95% confidence interval of $(0.69, 0.81)$. Note that the formulation of model (1) implicitly assumes that the response y_i represents the correct number of exacerbations for patient i. As mentioned above, however, the outcome in the analysis of Calverley *et al.*[19] stems from patient self-reports, thus such an assumption does generally not hold.

3 Analysis of validation data

To understand how reported and true values are related, *i. e.* to formulate an error model, it is crucial to rely on validation data. Ideally, such data stem from an internal source of information, for instance when the error-prone variable is measured according to some "gold standard" for a subset of the investigated

population. Given that no internal validation data were available in the context of the TORCH trial, we have extracted such relevant information from the ICE COLD ERIC study[20;21], where self-reported exacerbation numbers of 407 COPD patients were compared to the numbers ascertained by an adjudication committee who had access to the patients charts of their general practitioners, patient self-reports and from all follow-up assessments. Denote by $y_i^\star$ the self-reported number of exacerbations by patient i, while y_i is the corresponding true number of exacerbations. The aggregated validation data are shown in table 1. In a first attempt, we fitted a NB regression model

$$y_i^\star \,|\, y_i \sim \mathsf{NB}(\gamma_0 + \gamma_1 y_i, \theta_E)$$

with identity (id) link, regression parameters γ_0, γ_1 and overdispersion parameter θ_E to describe the distribution of the reported counts as a function of the true counts. To this end, a standard likelihood approach was used. The fact that the overdispersion parameter was estimated as $\hat{\theta}_E = 3.49 \ll \infty$ indicates that an error model with overdispersion is appropriate for the miscounting error in this study. Moreover, the estimated overdispersion parameter $\hat{\theta}_E$ was larger than when a negative binomial model with the more common log link, including $\log(y_i + 1)$ as explanatory variable, was used ($\hat{\theta} = 3.12$). Thus, the id link led to a model with less overdispersion, but also to a better model fit, as reflected by its AIC of 1271 compared to 1288 for the model with log link. However, deviance residuals indicated that there might be an excessive number of zeroes in the reported counts (figure 1, left). We therefore replaced the NB distribution of the error model by a ZINB distribution, leading to

$$y_i^\star \,|\, y_i \sim \mathsf{ZINB}\left(\gamma_0 + \gamma_1 y_i, p_i, \theta_E\right) \ , \tag{2}$$

with a parameterization as given in Appendix A. The zero-inflation probabilities p_i were related to y_i via the logistic transformation $\mathsf{logit}(p_i) = \delta_0 + \delta_1 \mathsf{I}(y_i > 0)$ with indicator covariate $\mathsf{I}(y_i > 0) = 1$ if $y_i > 0$ and 0 otherwise, which led to a better fit than when directly including y_i. Although it might not be directly evident from the deviance plot (figure 1, right), model (2) resulted in less overdispersion ($\hat{\theta} = 6.09$) and in an AIC that decreased considerably from 1271 to 1260.

Before combining the error model (2) with the regression model from the TORCH study, we discuss the novel modelling approach and the effect of miscounting error in outcomes in more detail, see the next section.

4 Modelling and effects of miscounted outcomes

4.1 Error modelling for count outcomes

A slightly more general formulation of the ZINB error model (2) is given by

$$y_i^\star \mid y_i \sim \mathsf{ZINB}(\mu_i, p_i, \theta_E) \ , \tag{3}$$

where the mean $\mu_i = h(\gamma_0 + \gamma_1 y_i)$ is linked to the true counts y_i via a possibly non-linear (inverse link) function h, and $\mathsf{logit}(p_i) = \boldsymbol{w}_i^\top \boldsymbol{\delta}$ depends on a vector $\boldsymbol{w}_i$ of covariates and a vector of regression coefficients $\boldsymbol{\delta}$. The id link is a natural choice when the true and the observed counts are on the same scale. Note that $\boldsymbol{w}_i$ can often simply be replaced by $(1, y_i)^\top$, but we used this more general notation here to allow for the inclusion of transformed versions of y_i and for additional covariates. If no excessive zeroes are expected, the ZINB regression can be replaced by the simpler NB model with $p_i = 0$. If in addition $\theta \to \infty$, the error distribution reduces to Poisson.

If the id link is used, the restrictions $\gamma_0 \geq 0$ and $\gamma_1 \geq 0$ avoid negative expected values and imply increasing mean and variance for larger true counts. In the special case of a NB error model with $\gamma_0 = 0$ and $\gamma_1 = 1$, model (3) implies unbiased error, *i. e.* $\mathsf{E}(y_i^\star \mid y_i) = y_i$. However, outcomes with $y_i = 0$ then necessarily lead to observations $y_i^\star = 0$, imposing that zero counts are always correctly reported. If such a restriction is unnatural, it can be avoided by requiring $\gamma_0 > 0$ to allow for over-reporting in the case $y_i = 0$. This consideration shows that unbiasedness is not an essential property of count error modelling, as the error distribution is not naturally symmetrical, in particular for small counts.

The use of a ZINB error model implies that the error variance is $\mathsf{Var}(y_i^\star \mid y_i) \geq \mathsf{E}(y_i^\star \mid y_i)$, and equality holds when the model is Poisson, *i. e.* $p_i = 0$ and $\theta_E = \infty$. The model thus imposes a minimal variance for the distribution of the observed counts around the true counts. In some situations such a modelling assumption could be implausible, in which case the ZINB error model may be replaced by a count model that allows for underdispersion, for instance the generalized event count model[28] or the generalized Poisson distribution[29]. However, overdispersion is not a critical assumption for the error in the response of the TORCH study that is analyzed here (figure 1).

The formulation of model (3) propagates a non-differential error, *i. e.* it implies that the error is independent of the covariates $(x_i, \boldsymbol{z}_i)$ given the true response y_i, and thus $\mathsf{Pr}(y_i^\star \mid y_i, x_i, z_i) = \mathsf{Pr}(y_i^\star \mid y_i)$. In a more general setup $y_i^\star$ may depend on the covariates $(x_i, \boldsymbol{z}_i)$, in which case the error in $y_i^\star$ is differential. To keep notation simple, however, we will in the following write the parametric models without covariate dependencies, except when explicitly needed.

4.2 The effect of a miscounted response

It is important to understand potential effects of error-prone count outcomes $y_i^\star$ on the parameter estimates in Poisson, NB or ZINB regression model. As discussed in section 4.1, error that is generated according to model (3) may induce bias in the observed counts $y_i^\star$, thus $\mathsf{E}(y_i^\star \mid y_i) \neq y_i$, and unbiased error is only retrieved in very special cases, *i. e.* when the model is NB with $h = \mathsf{id}$, $\gamma_0 = 0$ and $\gamma_1 = 1$. In this case, the parameter estimates for β_0 and β_x are unbiased in the naive regression as well, because $\log(\mathsf{E}(y_i^\star)) = \log(\mathsf{E}(y_i)) = \beta_0 + x_i\beta_x$. On the other hand, when $\gamma_0 = 0$ but $\gamma_1 \neq 1$, the error model is no longer unbiased, but when the standard log-link is used in the regression model (which is always the case here), the slope parameter β_x can still be consistently estimated, as can be seen from

$$\begin{aligned} \log(\mathsf{E}(\mathsf{E}(y_i^\star \mid y_i)) &= \log(\mathsf{E}(\gamma_1 y_i)) \\ &= \log(\gamma_1) + \log(\mathsf{E}(y_i)) \\ &= \log(\gamma_1) + \beta_0 + x_i\beta_x \ . \end{aligned} \tag{4}$$

Generally, the likelihood for an erroneous observed regression outcome $y_i^\star$ can be written as

$$\mathsf{Pr}(y_i^\star \mid x_i, \boldsymbol{z}_i) = \sum_{y_i} \mathsf{Pr}(y_i^\star \mid y_i, x_i, \boldsymbol{z}_i)\, \mathsf{Pr}(y_i \mid x_i, \boldsymbol{z}_i) \ . \tag{5}$$

If the error in $y_i^\star$ is non-differential, the expression $\mathsf{Pr}(y_i^\star \mid y_i, x_i, \boldsymbol{z}_i)$ can be replaced by $\mathsf{Pr}(y_i^\star \mid y_i)$. If there is no relationship between y_i and the covariates $(x_i, \boldsymbol{z}_i)$, both terms in (5) are independent of $(x_i, \boldsymbol{z}_i)$, and thus also $y_i^\star$ is independent of the predictors. A naive regression analysis then leads to valid conclusions about the association of the predictors with the true response, *i. e.* if the predictors

are independent of the response, non-differential error cannot induce a spurious effect. Still, the resulting tests have decreased power, as discussed by Carroll *et al.*[5][section 15.4]. Moreover, if the response and the predictors are associated, non-differential error typically leads to attenuated versions of the true effects, see *e. g.* Appendix B for the case that is relevant in this study, although situations with the opposite consequence, *i. e.* overestimation of the effect size, can be constructed (see the end of Appendix B for a hypothetical example).

On the other hand, if the error in $y_i^\star$ is differential, equation (5) shows that there may be a relationship between $y_i^\star$ and $(x_i, \boldsymbol{z}_i)$, even if the true response is not associated with the covariates. Classical hypothesis tests for the regression parameters β_x and $\boldsymbol{\beta}_z$ are then no longer valid and often lead to spurious significance and reverse attenuation. Finally, a true relation between the covariates and y_i may be masked by a non-differential *or* a differential error. Hence, the direction of a potential bias in the parameter estimates induced by the ME in $y_i^\star$ cannot be predicted in general.

4.3 Bayesian hierarchical model

Consider a regression model with count outcome y_i, potentially overdispersed and/or zero-inflated, and again assume that y_i can only be observed via a miscounted proxy $y_i^\star$. The error model (3) can then, in principle, be combined with any count regression model. Here, however, we assume that all extra-variability and zero-inflation in the measured response is attributed to the miscounting process. We therefore formulate a hierarchical model that comprises a Poisson

model for the true observations, and a ZINB error model:

$$y_i \sim \mathsf{Po}\left(\exp(o_i + \beta_0 + x_i\beta_x + \boldsymbol{z}_i\boldsymbol{\beta}_z)\right) , \quad (6)$$

$$y_i^\star \,|\, y_i \sim \mathsf{ZINB}\left(h(\gamma_0 + \gamma_1 y_i), p_i, \theta_E\right) , \quad (7)$$

where $o_i = \log(t_i)$ denotes the offset, $\mathsf{logit}(p_i) = \boldsymbol{w}_i^\top \boldsymbol{\delta}$, and h is again the inverse link function of the error model. The Poisson assumption for the regression model involving the true counts could easily be relaxed by using a NB or ZINB model, but it may then be difficult to identify the various contributors to the variance of $y_i^\star$, or to separate the zero-inflation mechanisms of the regression and the error model, in particular if only weak prior information is available. Still, such an approach may be sensible, for instance when there are specific reasons to expect zero-inflation in the regression model. With respect to the estimates of the regression model parameters, the choice is not expected to be critical, which is also confirmed by additional calculations presented in the supplementary material (for results see Table S1). Independent normal priors with small precision are usually specified for β_0, β_x and the components of $\boldsymbol{\beta}_z$. Information on the parameters of the error model, namely $(\gamma_0, \gamma_1, \boldsymbol{\delta}^\top)^\top$ and θ_E, must be obtained from (internal or external) validation data or expert knowledge. If the error model is expected to be covariate-dependent, it is beneficial if the model parameters can be estimated from separate validation data (sub)sets.

The marginal distribution of the measured response $y_i{}^\star$ following model (6)-(7) is overdispersed by construction. This also holds if $y_i^\star \,|\, y_i$ is Poisson distributed,

i. e. when $\theta_E = \infty$, in which case the marginal expectation and variance of $y_i^\star$ are

$$\begin{aligned} \mathsf{E}(y_i^\star) &= \mathsf{E}(\mathsf{E}(y_i^\star \mid y_i)) \\ \mathsf{Var}(y_i^\star) &= \mathsf{E}(\mathsf{Var}(y_i^\star \mid y_i)) + \mathsf{Var}(\mathsf{E}(y_i^\star \mid y_i)) \\ &= \mathsf{E}(\mathsf{E}(y_i^\star \mid y_i)) + \mathsf{Var}(\mathsf{E}(y_i^\star \mid y_i)) \ . \end{aligned}$$

The last equality holds because variance and expected value are equal under the Poisson assumption. Thus in general we have $\mathsf{Var}(y_i^\star) > \mathsf{E}(y_i^\star)$, *i. e.* overdispersion. Therefore, our proposed error modelling framework should only be applied if the observed counts are (marginally) overdispersed.

A general concern in ME modelling is the aspect of identifiability, namely when the error model parameters are unknown[30]. Equation (4), for instance, illustrates that confounding between γ_1 and β_0 could be an issue. Interestingly, however, Gustafson[30] has illustrated that already relatively crude priors can be sufficient to obtain good results if there is enough indirect learning about nonidentifiable model parameters.

Marginal posterior distributions for the parameters of model (6)-(7) can be obtained by MCMC sampling. A simulation example including code is given as online supplementary material (files 2 and 3). Unfortunately, as the latent variable $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$ is not Gaussian, it is not possible to approximate the posterior marginals by integrated nested Laplace approximations (INLA), which are a computationally convenient alternative to sampling approaches for Bayesian inference in latent Gaussian models[31], in particular in the presence of covariate measurement error[32].

5 Application to the TORCH study

5.1 Non-differential response error modelling

In this section, we use the prior information derived from the ICE COLD ERIC study to reanalyze the TORCH data. We thus eventually apply the modelling framework as given in (6)-(7) by combining the ZINB error model with id link and a Poisson regression model to obtain corrected estimates for the effect β_x of the combination treatment in the TORCH study. The hierarchical model now is

$$\begin{aligned} y_i &\sim \mathsf{Po}\left(\exp(\log(t_i) + \beta_0 + x_i\beta_x + \boldsymbol{z}_i\boldsymbol{\beta}_z)\right) \ , \\ y_i^\star \,|\, y_i &\sim \mathsf{ZINB}\left(\gamma_0 + \gamma_1 y_i, p_i, \theta_E\right) \ , \end{aligned}$$

with $\mathsf{logit}(p_i) = \delta_0 + \delta_1 \mathsf{I}(y_i > 0)$. The model thus essentially estimates only two distinct zero-inflation probabilities, namely one for individuals with $y_i = 0$ (*i. e.* those patients that had no exacerbation), and one for those with $y_i > 0$.

Prior information on the error model parameters $\boldsymbol{\alpha} = (\gamma_0, \gamma_1, \delta_0, \delta_1)$ and θ_E was obtained as described in Section 3, *i. e.* by fitting model (2) to our validation data without stratification for treatment or any additional covariates. The miscounting error in the reported exacerbation counts was thus assumed to be non-differential in this first error model. Maximum-likelihood estimates of the error model parameters $\boldsymbol{\alpha}$ and the corresponding covariance matrix $\boldsymbol{\Sigma}$ were given as

$$\hat{\boldsymbol{\alpha}} = (0.753, 0.966, 0.151, -3.174)^\top \ , \tag{8}$$

$$\hat{\boldsymbol{\Sigma}} = \begin{pmatrix} 0.020 & -0.007 & 0.033 & -0.019 \\ -0.007 & 0.007 & -0.011 & 0.018 \\ 0.033 & -0.011 & 0.122 & -0.094 \\ -0.019 & 0.018 & -0.094 & 0.401 \end{pmatrix} . \tag{9}$$

It would be tempting to use Gaussian priors for $\boldsymbol{\alpha}$ with moments as estimated in (8) and (9). However, by doing this we would postulate that the error model deduced from the ICE COLD ERIC study is *transportable* to the TORCH study, *i. e.* that the mechanisms inducing the misreporting are exactly the same in both datasets. At first glance this seems to be a plausible assumption as both datasets were collected over the same duration of three years, and the mean number of reported exacerbations were similar (2.02 in the validation data and 2.19 in the TORCH study). However, such an assumption can still lead to a prior-data conflict, indicated for instance by a low Box's p-value[25], particularly if the conditions under which the validation data were collected differ from the study conditions, which is difficult to verify. In our example, the operationalization of exacerbation measurements might deviate among the studies, or the ensembles could encompass a different mixture of ethnicities or disease severities. We therefore used a recently suggested approach by Held and Sauter[26], termed *adaptive prior weighting*, which at the same time identifies and accounts for a potential prior-data conflict. The idea is to multiply the covariance matrix from the validation data with an unknown scalar $g > 0$, leading to the prior

$$\boldsymbol{\alpha} \,|\, g \sim \mathsf{N}(\hat{\boldsymbol{\alpha}}, g\hat{\boldsymbol{\Sigma}}) \;, \tag{10}$$

with a uniform prior for

$$\frac{g}{g+1} \sim \mathrm{U}(0,1) \;, \tag{11}$$

i. e. a hyper-g prior[33;34] for g. This design allows to adaptively weight the prior distribution with weight $w = 1/g$. Values of $w < 1$ (*i. e.* $g > 1$) then indicate that the prior distribution is downweighted due to a prior-data conflict, and the prior distribution becomes flatter. On the other hand, values $w > 1$ (*i. e.* $g < 1$) increase

the weight of the prior by narrowing the prior distribution, which suggests that the prior is in good agreement with the data[26]. The overdispersion parameter θ_E was estimated from the validation data as $\hat{\theta}_E = 6.09$ with a standard error of 2.03. A log-normal prior distribution $\theta_E \sim \mathsf{LN}(\log(6.09), 0.33^2)$ was therefore used, where the second argument is the squared standard error, calculated using the delta-rule: $\mathsf{se}(\log(\hat{\theta}_E)) = 2.03/6.09 = 0.33$. Finally, independent $\mathsf{N}(0, 10^2)$ priors were assigned to the components of $\boldsymbol{\beta} = (\beta_0, \beta_x, \boldsymbol{\beta}_z^\top)^\top$.

A Bayesian analysis using MCMC was performed in JAGS via the `R`-interface `rjags`[35;36] by running two parallel chains for 25 000 iterations each, with a burn-in of 2 500 and a saving frequency of 5, and both chains were used for estimation. The posterior mean of the rate ratio for treatment vs. placebo was $\exp(\hat{\beta}_x) = 0.80$ with a 95% credible interval (CI) ranging from 0.72 to 0.89. The graph labelled as Corrected in figure 2a) depicts this estimate in comparison to the uncorrected estimate $\exp(\hat{\beta}_x) = 0.75$ with 95% confidence interval $(0.69, 0.81)$ from Calverley *et al.*[19]. Error-correction hence led to an estimate closer to 1, *i. e.* to a less pronounced treatment effect. Analytical considerations however show that non-differential response error in the model used here necessarily leads to attenuation effects, see Appendix B, while a correction towards 1 would imply that the error caused reverse attenuation. Thus, if the error model was specified correctly, we would expect a correction in the opposite direction of what is observed here, indicating that either the model did not correctly capture the error structure, or that there is an error in the model formulation. This concern is also supported by the posterior of g with a median of 33.3 (95% CI: 4.76,250), meaning that the weight of the prior was substantially decreased by a median factor of $\hat{w} = 1/\hat{g} = 0.03$ (95% CI: 0.004 to 0.21). Even without prior weighting ($g = 1$) the treatment effect was estimated as 0.79 (95% CI: 0.71 to 0.88), and

also the assignment of a fixed prior to $\boldsymbol{\alpha}$, *i. e.* by setting $g = 0$, did not change the quality of the result (estimate 0.78, 95% CI: 0.71 to 0.87).

In order to better understand the TORCH data structure, the same analysis was carried out separately on three subgroups of patients. The first group consisted of the 1973 patients that were under treatment for the duration of at least 2.5 years (the maximum duration was 3.1 years). The naive analysis for this group resulted in an estimated treatment effect of $\exp(\hat{\beta}_x) = 0.89$ (95% CI: 0.72 to 1.10). MCMC was then used to fit the error model with the same model parameters as in equations (8) and (9) and prior weighting according to (10) and (11), which yielded an estimate of $\exp(\hat{\beta}_x) = 0.82$ (95% CI: 0.70 to 0.94). The same direction of the correction was observed for the second group of 522 patients with a treatment time between 1 and 2.5 years (naive estimate of $\exp(\hat{\beta}_x) = 0.80$, 95% CI: 0.63 to 0.92 vs. error-corrected estimate of 0.70, 95% CI: 0.51 to 0.92), and for the third group of 653 patients that were under treatment for ≤ 1 year (naive and error-corrected estimates equal to 0.58, 95% CI from 0.48 to 0.71, and 0.52, 95% CI from 0.34 to 0.74, respectively). All estimates and their uncertainties are shown in figures 2b)-d). The results illustrate two things: First, the observed treatment effect gradually weakens over time. And second, the effect becomes stronger after error-correction within the three subgroups. This indicates that the above finding, where the treatment effect weakened upon error correction, was indeed the consequence of a model misspecification. It is not straightforward to isolate the origin of the observed time-dependency in the treatment effect. A likely explanation is that the data suffer from a so-called *emigrative selection bias*[37], which emerges when withdrawing rates in the placebo and the treatment groups

differ, and when at the same time withdrawing patients tend to have more severe disease, which was exactly the case in the TORCH trial[19;38].

Given that the scope of the present paper is to address miscounting error and not that of selection bias, a simple and pragmatic way to capture the time-dependency of the treatment effect is to include an interaction between the treatment x_i of patient i and the (log-transformed) time under treatment t_i. Therefore, we extended the original model (1) that was used in the analysis of the TORCH trial by an interaction term $x_i \log(t_i)$ and a main effect $\log(t_i)$:

$$y_i \sim \mathsf{Po}\left(\exp(\log(t_i) + \beta_0 + x_i\beta_x + x_i \log(t_i)\beta_{xt} + \log(t_i)\beta_t + \boldsymbol{z}_i\boldsymbol{\beta}_z)\right) \ , \qquad (12)$$

where β_{xt} and β_t are the respective slope parameters. Importantly, the ΔAIC between the original model (1) and the extended model (12) is -275 for the likelihood analysis without error modelling, indicating a very clear improvement of the model fit. Model (12) was then again enhanced with the error model as specified in equations (8)-(11), and the posterior distribution was estimated via MCMC sampling, using the same length and number of chains. The treatment and interaction effects of model (12) without error modelling were then given as $\exp(\hat{\beta}_x) = 0.74$ (95% CI: 0.66 to 0.83) and $\exp(\hat{\beta}_{xt}) = 1.14$ (95% CI: 1.03 to 1.27), while error correction lead to $\exp(\hat{\beta}_x) = 0.70$ (95% CI: 0.57 to 0.86), with $\exp(\hat{\beta}_{xt}) = 1.18$ (95% CI: 0.97 to 1.44). This means, as an example, that for a treatment time of $t_i = 1$ (in years), the treatment effect is corrected from 0.74 to 0.70, while for $t_i = 3$, the effect changes from 0.86 to 0.84. Overall, error correction was now in the expected direction, meaning that the treatment effect is underestimated when miscounted outcomes are used without any correction. Interestingly, the regression parameter β_t was negative, both in the model with and without error considerations ($\hat{\beta}_t = -0.50$, 95% CI from -0.56 to -0.45 without

error modelling, and $\hat{\beta}_t = -0.33$, 95% CI from -0.45 to -0.19 with error modelling), which reflects that patients with severe disease were more likely to withdraw from the study[38]. The posterior median of the prior weight $\hat{w}$ of the error model parameters was now 0.07 (95% CI from 0.01 to 0.34), thus slightly larger than when using the regression model without interaction term, although there is still indication for a prior-data conflict that cannot be resolved with the available datasets.

5.2 Differential response error modelling

Interestingly, our validation dataset provided information on the sex of the patients, and it was thus possible to estimate the error model components separately for females ($n_1 = 176$ patients) and males ($n_2 = 231$ patients). The respective data are given in tables S2 and S3 of the supplementary pdf file. The sex-specific parameter estimates and covariance matrices were then

$$\hat{\boldsymbol{\alpha}}^{(1)} = (0.688, 1.064, -0.356, -12.811)\ , \tag{13}$$

$$\hat{\boldsymbol{\Sigma}}^{(1)} = \begin{pmatrix} 0.041 & -0.017 & 0.097 & -0.006 \\ -0.017 & 0.017 & -0.040 & 0.120 \\ 0.097 & -0.040 & 0.470 & -0.180 \\ -0.006 & 0.120 & -0.180 & 55370 \end{pmatrix}\ , \tag{14}$$

for females and

$$\hat{\boldsymbol{\alpha}}^{(2)} = (0.824, 0.878, 0.502, -2.799) \ , \tag{15}$$

$$\hat{\boldsymbol{\Sigma}}^{(2)} = \begin{pmatrix} 0.036 & -0.012 & 0.046 & -0.030 \\ -0.012 & 0.009 & -0.014 & 0.017 \\ 0.046 & -0.014 & 0.155 & -0.132 \\ -0.030 & 0.017 & -0.132 & 0.308 \end{pmatrix} \tag{16}$$

for males, with overdispersion parameters estimated as $\hat{\theta}_E^{(1)} = 4.13\,(\mathsf{se} = 1.33)$ and $\hat{\theta}_E^{(2)} = 14.71\,(\mathsf{se} = 12.18)$. Note that the large variance for the parameter δ_1 in the female group (the entry in the lower right corner of matrix (14)) indicates that the respective parameter is essentially nonidentifiable. This problem could arise because the zero-inflation probability for females with true exacerbations $y_i > 0$ was essentially zero, *i. e.* females did then not report excessive zeroes, so that $\delta_1^{(1)}$ becomes small, and thus difficult to be estimated. An additional model checking step then revealed that the error distribution for the female group can be described by the simpler NB model, *i. e.* ignoring zero-inflation (AIC for NB: 591, AIC for ZINB: 592), while the ZINB model is needed for the males (AIC improvement from 681 to 669 when changing from NB to ZINB). The ZINB error model (13)-(14) for females was thus replaced by the simpler NB model, (which corresponds to $p_i = 0$ for female patients), and the remaining parameters were estimated as

$$\hat{\boldsymbol{\alpha}}^{(1)} = (\hat{\gamma}_0^{(1)}, \hat{\gamma}_1^{(1)}) = (0.421, 1.183) \ , \quad \hat{\boldsymbol{\Sigma}}^{(1)} = \begin{pmatrix} 0.007 & -0.003 \\ -0.003 & 0.012 \end{pmatrix} \ , \tag{17}$$

and $\hat{\theta}_E^{(1)} = 3.74\,(\mathsf{se} = 1.16)$. The sex-dependent error in the response was then modelled using the estimates and covariance matrices as given in (15),

(16) and (17), with log-normal priors $\theta_E^{(1)} \sim \mathsf{LN}(\log(3.74), 0.31^2)$ and $\theta_E^{(2)} \sim \mathsf{LN}(\log(14.71), 0.83^2)$, again applying the delta-rule to obtain the variances. As before, the covariance matrices $\mathbf{\Sigma}^{(1)}$ and $\mathbf{\Sigma}^{(2)}$ were adaptively weighted by scalars g_1 and g_2, which were given hyper-g priors as in (11) in order to account for a potential prior-data conflict.

Posterior distributions were again estimated with MCMC using the same setup as above, but running the chains for 100 000 iterations each to ensure convergence. The posterior means for the treatment and interaction effects were given as $\exp(\hat{\beta}_x) = 0.69$ (95%-CI: 0.56 to 0.84) and $\exp(\hat{\beta}_{xt}) = 1.19$ (95% CI: 0.98 to 1.45), respectively. Median prior weights for the two groups were estimated as $\hat{w}_1 = 0.50$ (95% CI: 0.03 to 15.26) and $\hat{w}_2 = 0.06$ (95% CI: 0.01 to 0.30), still indicating some prior-data conflict in the male subgroup. Interestingly, the switch from a non-differential to a covariate-dependent response error model did essentially not change the results, thus a non-differential model would be sufficient here. Still, the exercise illustrates that covariate-dependent error modelling is conceptually straightforward, given that prior information is available.

6 Simulation study

The error analysis of the TORCH trial did not only reveal a time-dependency of the treatment effect of the originally presented regression model, but also illustrated that error-modelling on top of a misspecified regression model may lead to corrections of effect estimates that are in an unexpected or even wrong direction. It is useful to perform three simulations to illustrate that (i) error modelling as proposed here leads to unbiased estimates of the true effects, given

that all modelling assumptions are fulfilled, and that (ii) the effects observed in the error analysis of the TORCH study may indeed originate from a model misspecification, such as an omitted time-dependency of the treatment effect. In each example, $n = 2000$ count outcomes y_i were generated in dependence of a treatment indicator $\boldsymbol{x}$ with $x_i = 0$ for $i = 1, \ldots, 1000$ and $x_i = 1$ otherwise, and an offset $\log(t_i)$ such that t_i was uniformly distributed between 0.1 and 3. In all simulations, miscounting error in y_i was generated according to the ZINB error model

$$y_i^\star \,|\, y_i \quad \sim \quad \mathsf{ZINB}\left(\gamma_0 + \gamma_1 y_i, p_i, \theta_E\right) \tag{18}$$

with $\mathsf{logit}(p_i) = \delta_0 + \delta_1 \mathsf{I}(y_i > 0)$ and error model parameters $(\gamma_0, \gamma_1, \delta_0, \delta_1) = (0.2, 1.2, 0.15, -3)$ and $\theta = 4$, which were chosen to be comparable to the values observed in the TORCH study. Each simulation was repeated 250 times. In each iteration, parameter estimates for the treatment effect β_x from the following three models were stored:

(i) the maximum likelihood (ML) estimate including the data without error in the response, leading to the 'error-free' estimate.

(ii) the ML estimate using the data with miscounting error in the response, leading to the 'naive' estimate.

(iii) the posterior mean of an MCMC sample for the respective Bayesian hierarchical error model with a burn-in of 1000 and a sampling of 5 000 iterations, leading to the 'corrected' estimate. To this end, the data-generating error model (18) was used with point priors for the error model parameters $(\gamma_0, \gamma_1, \delta_0, \delta_1) = (0.2, 1.2, 0.15, -3)$, and log-normal priors $\theta_E \sim$

$\mathsf{LN}(4, 1)$. Independent zero-mean Gaussian priors with a variance of 10^2 were specified for the slope parameters of the regression.

ML estimates were obtained using the `glm.nb()` and `glm()` functions in `R`, while MCMC samples were generated in `rjags`. ML estimates or posterior means with 2.5% and 97.5% quantile intervals from the 250 iterations were then plotted in figure 4.

Simulation 1: Simple regression model

In the first example, the regression model was given as

$$y_i \sim \mathsf{Po}\left(\exp(\log(t_i) + \beta_0 + x_i\beta_x)\right) ,$$

with regression model parameters $(\beta_0, \beta_x) = (1, \log(0.7))$. The estimates from analyses (i)-(iii) were stored in each iteration and displayed as boxplot representations in figure 4, left. The results illustrate that miscounting error leads to an attenuated version of the estimated treatment effect, but that the hierarchical error model retrieves unbiased estimates of the true effect.

Simulation 2: Time-dependent treatment effect, wrong model

In this second example, the true counts y_i were generated according to the regression model

$$y_i \sim \mathsf{Po}\left(\exp(\log(t_i) + \beta_0 + x_i\beta_x + \log(t_i)x_i\beta_{xt})\right) \quad (19)$$

with parameters $(\beta_0, \beta_x, \beta_{xt}) = (1, \log(0.7), 0.2)$. Analyses (i)-(iii) were then carried out, however choosing the regression model that did not include the interaction term $\log(t_i)x_i$, that is, β_{xt} was (erroneously) set to zero, while the

offset $\log(t_i)$ was included correctly. Please note that here (ii) and (iii) correspond to the original analysis that was carried out in the TORCH study without error modelling, and to the error-correction approach described in Section 6 where no time-dependency was included in the regression model. The results in the middle panel of figure 4 confirm the pattern that was observed in Section 6: The naive analysis leads to treatment effects that are stronger than when correct responses are included in the regression model. Moreover, error-correction leads to weaker overall effect estimates, that is, a correction towards the Null.

Simulation 3: Time-dependent treatment effect, correct model

In this last case data were again generated according to model (19), but this time the correct regression model was included in all analyses. The regression model (19) was then used (i) once with correct y_i, (ii) once with naive response $y_i^\star$, and (iii) once for the respective hierarchical error model. Cases (ii) and (iii) correspond to the naive and error-corrected versions of the TORCH analysis when the interaction term $x_i \log(t_i)$ was added to the model. The results in the right panel of figure 4 confirm that non-differential response error leads to an attenuated treatment effect estimate, and that error analysis is able to properly correct for it.

7 Discussion

We have proposed a novel statistical framework to treat error in count outcomes by formulating a ZINB error model. This parametric model only requires prior elicitation on a few model parameters, and does not artificially limit the range of the true counts, in contrast to previously suggested methodology[22]. We have shown how a Bayesian hierarchical model, including a Poisson regression

model and a ZINB error model, can be employed to jointly estimate posterior marginals via MCMC sampling. The development of the methodology proposed here was motivated by the TORCH trial, where the efficacy of a treatment on the exacerbation rate of COPD patients was studied by regressing the exacerbation numbers from patient self-reports on the covariates using a NB model[19]. In an external study, these self-reported values have recently been shown to suffer from considerable miscounting error[20;21], thus the respective data could be used to formulate a count error model. A ZINB regression model of observed vs. true counts captured the error in the self-reported values reasonably well (figure 1), and informed priors for the respective model parameters could be extracted at the same time. Note that the model imposes a minimal error variance $\mathsf{Var}(y_i^\star \mid y_i) \geq \mathsf{E}(y_i^\star \mid y_i)$, which in fact appears to be suitable in our case, but may sometimes be inappropriate. Underdispersed count models might then be a solution, although we have not discussed them here.

When accounting for error in the outcome of the TORCH study using the proposed error model, the corrected treatment effect became weaker, indicating that it had been overestimated in the original study. Only thanks to closer inspection of the original regression model it became evident that this correction was an artefact of an unaccounted time-dependency of the treatment effect. This time-dependency, in turn, could itself be an artefact of non-random withdrawings (patients with severe disease withdrawings more often), which are known to potentially lead to biased estimates of model parameters. Accounting for such an emigrative selection bias would require a thorough understanding of withdrawing patterns, as well as additional modelling steps, but this is not the scope of this work. Therefore, the time-dependency was captured by a simple interaction term. Using this adapted model, error correction led to an increased effect size

estimate. This example clearly illustrates that error modelling is sensitive to model misspecifications, and that unexpected bias corrections may therefore help to discover relevant problems.

Most applied researchers are aware of biases induced by ME in covariates or in the response of regression models, it is often assumed that the observed effects are then conservative estimates of the true effects. However, this implicitly postulates certain types of error structures in the observed variables, typically non-differential ME. Although this assumption seems to hold in the case study presented here, it may be violated more often than believed. In fact differential error sometimes emerges unexpectedly. As an example, Mwalili *et al.*[22] have shown how the combination of miscounted values from several examiners in an oral health study leads to differential error globally, even if the miscounting process of each examiner is non-differential. Similarly, Gustafson[7] illustrated how the categorization of a continuous covariate suffering from non-differential ME can induce a differential misclassification error. The fact that such an error may lead to nonconservative estimated effects in clinical or epidemiological studies is critical and should not be ignored. An example of how to capture covariate-dependent ME was presented in Section 5.2.

A necessary prerequisite for any error modelling attempt is the availability of validation data. Here, we have emphasized that the elicitation of suitable priors for error model parameters can be challenging, and that priors derived from external validation data may introduce some prior-data conflict into the model of interest. An implicit assumption in the context of prior information transfer typically is that such validation data are transportable among studies, *i. e.* that the circumstances under which the validation and the study data were collected are comparable, as the information in the validation data does otherwise not lead

to sensible priors for the study data analysis. In our example the validation and study data were collected over the same duration (three years), and the reported exacerbation numbers were comparable. Still, it will usually be impossible to check such transportability premises. We have therefore suggested to adaptively weight the error model priors using a hyper-g prior to ensure that they are automatically downweighted in the presence of a prior-data conflict, *i. e.* when the transportability assumption is questionable. In this case study exactly such a down-weighting effect was observed, and we can only hypothesize why this was the case. Possible explanations could be differences in the composition of the study ensembles regarding, *e. g.* cultural or health state, or distinct standard operating procedures to assess exacerbations. Importantly, such transportability concerns could be mostly eliminated if internal validation data were available.

In conclusion, we have discussed that error in count outcomes may bias parameter estimates of regression models, and that the bias may be in any direction. The importance and also some difficulties of error modelling were highlighted, particularly in the context of clinical trials, where a crucial assumption is that effect estimates originate from conservative estimation procedures. We have introduced a parametric miscounting error modelling framework that is able to treat unbounded counts and seems to capture the error mechanism in the miscounted outcome of our case study reasonably well. Advantages and limitations of this novel approach were discussed, and in particular we recommend to check whether (or to justify why) the ZINB error model gives a realistic description of the miscounting process under consideration. Nevertheless, probably the best way to circumvent expensive and tricky error modelling procedures is to directly optimize the quality of the data. It is not surprising that Breslow (2014) wrote[39]: "Obviously, [...] the best method of dealing

with measurement error was to avoid it!" In the example of the TORCH study this could have been achieved by replacing patient self-reports by ascertained values obtained from an adjudication committee. Although this will obviously lead to higher costs per patient, such an extra effort may be worthwhile: Not only are more valid effect estimates expected, but also smaller sampling sizes might be sufficient thanks to the removal of uncertainty (*i. e.* error), leading to lower overall costs.

Acknowledgements

The TORCH study was sponsored by GlaxoSmithKline Research and Development Ltd, Brentford, Middlesex. We thank the study sponsor for access to the data. We especially thank Anja Frei who provided us with the validation dataset and stratified subsets to estimate the miscounting error model(s). We also thank Lukas Keller from the University of Zurich, who stimulated our work on measurement error modelling. This work would not have been possible without generous funding from the Faculty of Science of the University of Zurich.

Declaration of conflicting interests

The Authors declare that there is no conflict of interest.

References

1. Pearson K. On the mathematical theory of errors of judgement. *Philosophical Transactions of the Royal Society of London A* 1902; 198: 235–299.
2. Wald A. The fitting of straight lines if both variables are subject to error. *Annals of Mathematical Statistics* 1940; 11: 284–300.

3. Berkson J. Are there two regressions? *Journal of the American Statistical Association* 1950; 45: 164–180.
4. Fuller WA. *Measurement Error Models*. New York: John Wiley & Sons, 1987.
5. Carroll RJ, Ruppert D, Stefanski LA et al. *Measurement Error in Nonlinear Models: A Modern Perspective*. 2nd ed. Boca Raton: Chapman & Hall, 2006.
6. Buonaccorsi JP. *Measurement Error: Models, Methods, and Applications*. Boca Raton, FL: CRC Press, 2010.
7. Gustafson P. *Measurement Error and Misclassification in Statistics and Epidemiology: Impacts and Bayesian Adjustments*. Boca Raton: Chapman & Hall/CRC, 2004.
8. Abbrevaya J and Hausman JA. Response error in a transformation model with application to earnings-equation estimation. *Econometrics Journal* 2004; 7: 366–388.
9. Carroll RJ and Ruppert D. *Transformation and Weighting in Regression*. London: Chapman & Hall, 1988.
10. Buonaccorsi JP. Measurement error, linear calibration and inferences for means. *Computational Statistics and Data Analysis* 1991; 11: 239 – 257.
11. Buonaccorsi JP and Tosteson T. Correcting for nonlinear measurement error in the dependent variable in the general linear model. *Communications in Statistics, Theory & Methods* 1993; 22: 2687 – 2702.
12. Buonaccorsi JP. Measurement error in the response in the general linear model. *Journal of the American Statistical Association* 1996; 91: 633 – 642.
13. Ekholm A and Palmgren J. Correction for misclassification using doubly sampled data. *Journal of Official Statistics* 1987; 3: 419–429.
14. Copas JB. Binary regression models for contaminated data (with discussion). *Journal of the Royal Statistical Society Series B (Statistical Methodology)* 1988; 50:

225–265.

15. Neuhaus JM. Bias and efficiency loss due to misclassified responses in binary regression. *Biometrika* 1999; 86: 843–855.
16. Neuhaus JM. Analysis of clustered and longitudinal binary data subject to response misclassification. *Biometrics* 2002; 58: 675–683.
17. Magder LS and Hughes JP. Logistic regression when the outcome is measured with uncertainty. *American Journal of Epidemiology* 1997; 146: 195–203.
18. Paulino CD, Soares P and Neuhaus J. Binomial regression with misclassification. *Biometrics* 2003; 59: 670–675.
19. Calverley PM, Anderson JA, Celli B et al. Salmeterol and fluticasone propionate and survival in chronic obstructive pulmonary disease. *New England Journal of Medicine* 2007; 356: 775–789.
20. Siebeling L, Puhan MA, Muggensturm P et al. Characteristics of Dutch and Swiss primary care COPD patients – baseline data of the ICE COLD ERIC study. *Clinical Epidemiology* 2011; 3: 273–283.
21. Frei A, Siebeling L, Wolters C et al. The inaccuracy of patient recall for COPD exacerbation rate estimation and its implications: Results from central adjudication. *CHEST* 2016; 150: 860–868.
22. Mwalili S, Lesaffre E and Declerck D. The zero-inflated negative binomial regression model with correction for misclassification: An example in caries research. *Statistical Methods in Medical Research* 2008; 17: 123–139.
23. Albert P, Hunsberger S and Biro F. Modeling repeated measures with monotonic ordinal responses and misclassification, with applications to studying maturation. *Journal of the American Statistical Association* 1997; 92: 1304–1311.
24. Prentice RL. Surrogate endpoints in clinical trials: Definition and operational criteria. *Statistics in Medicine* 1989; 8: 431–440.

25. Box GEP. Sampling and Bayes' inference in scientific modelling and roubstness. *Journal of the Royal Statistical Society Series A (Statistics in Society)* 1980; 143: 383–430.

26. Held L and Sauter R. Adaptive prior weighting in generalized linear models. *Biometrics* 2016; Early View at `http://dx.doi.org/10.1111/biom.12541`.

27. Suissa S. Statistical treatment of exacerbations in therapeutic trials of chronic obstructive pulmonary disease. *American Journal of Respiratory and Critical Care Medicine* 2006; 173: 842–846.

28. King G. Variance specification in event count models: From restrictive assumptions to a generalized estimator. *American Journal of Political Science* 1989; 33: 762–784.

29. Consul PC and Jain GC. A generalization of the Poisson distribution. *Technometrics* 1973; 15: 791–799.

30. Gustafson P. On model expansion, model contraction, identifiability and prior information: Two illustrative scenarios involving mismeasured variables. *Statistical Science* 2005; 20: 111–140.

31. Rue H, Martino S and Chopin N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations (with discussion). *Journal of the Royal Statistical Society Series B (Statistical Methodology)* 2009; 71: 319–392.

32. Muff S, Riebler A, Held L et al. Bayesian analysis of measurement error models using integrated nested Laplace approximations. *Journal of the Royal Statistical Society Series C (Applied Statistics)* 2015; 64: 231–252.

33. Liang F, Paulo R, Molina G et al. Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association* 2008; 103: 410–423.

34. Held L, Sabanés Bové D and Gravestock I. Approximate Bayesian model selection with the deviance statistic. *Statistical Science* 2015; 30: 242–257.

35. Plummer M. JAGS: A Program for Analysis of Bayesian Graphical Models Using Gibbs Sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*.

36. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013. URL `http:/f/www.R-project.org/`.

37. Mansournia MA, Hernán MA and Greenland S. Matched designs and causal diagrams. *International Journal of Epidemiology* 2013; 42: 860–869.

38. Keene ON, Vestbo J, Anderson JA et al. Methods for therapeutic trials in COPD: lessons from the TORCH trial. *European Respiratory Journal* 2009; 34: 1018–1023.

39. Breslow NE. Lessons in biostatistics. In Lin X, Genest C, Banks DL et al. (eds.) *Past, Present and Future of Statistical Science*. CRC Press, 2014. pp. 335–347.

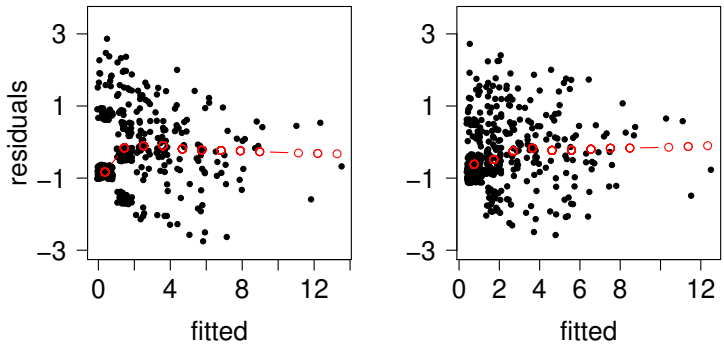

Figure 1. Deviance residuals versus fitted values of the negative binomial (left) and the zero-inflated negative binomial regression (right) using the validation dataset, where the centrally adjudicated exacerbation counts were fitted against the respective patient self reports. A small jitter has been added to both the fitted values and the residuals. The red line gives a standard lowess smoother and shows some evidence for zero-inflation.

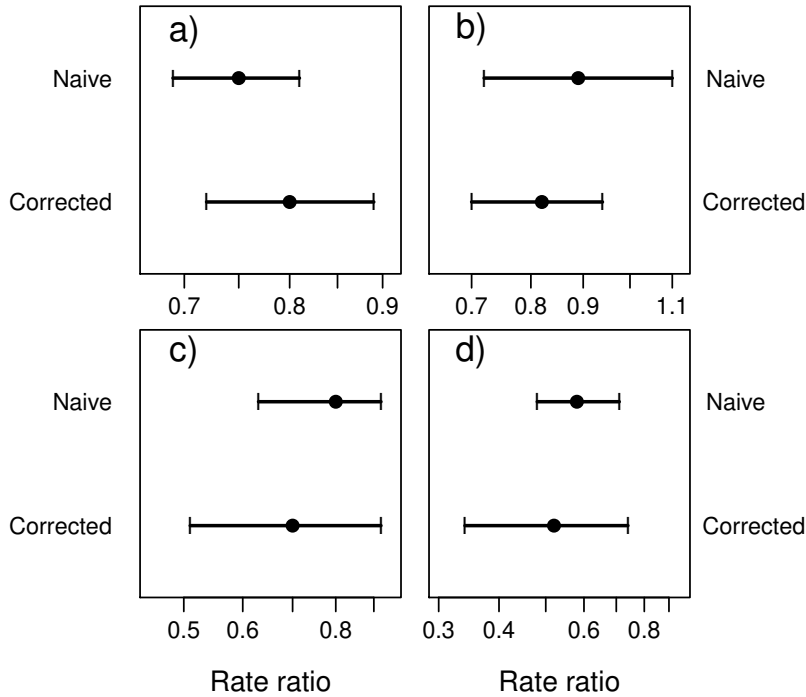

Figure 2. Naive and error-corrected estimates for the treatment vs. placebo rate ratio $\exp(\beta_x)$ of exacerbation rates in the TORCH trial. The horizontal lines represent 95% confidence/credible intervals. The x-axis is given in log-scale. a) "Corrected" shows the result for the case when non-differential error modelling was carried out on top of the original regression model used in the TORCH study. Panels b)-d) show the respective results for the subset of patients with a treatment time of > 2.5 years, between 1 and 2.5 years, and ≤ 1 year, respectively.

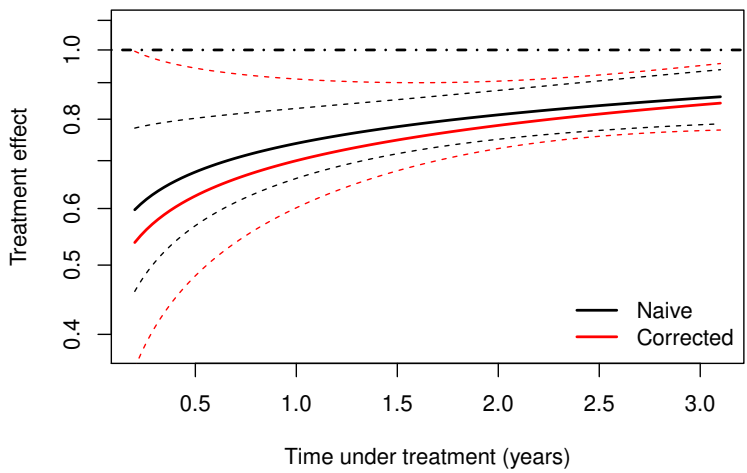

Figure 3. Time-dependent effect estimates of exacerbation rates for naive and error-corrected treatment vs. placebo rate ratios, calculated as $\exp(\hat{\beta}_x + \hat{\beta}_{xt} \log(t))$. Dashed lines indicate pointwise 95% CIs, and the black dash-dotted line refers to a rate ratio of 1 (no effect). The y-axis is given in $\log$ scale.

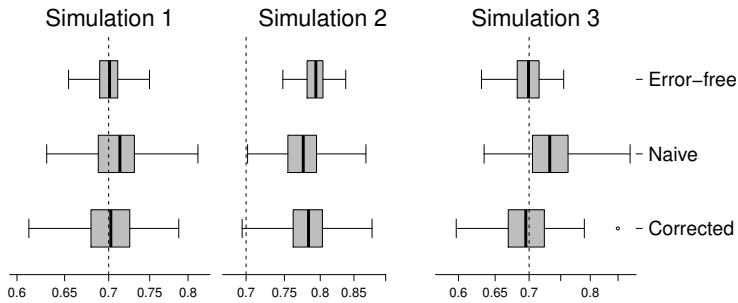

Figure 4. Boxplots for the ML estimates of error-free and naive estimates, as well as for the posterior means for the error-corrected estimates of the treatment vs. placebo rate ratio $\exp(\beta_x)$. Each boxplot was generated from the 250 iterations of the simulations. Note that the x-axis is given in log-scale.

	0	1	2	3	4	5	6	7	8	9	10	11	12
0	127	24	5	4	2	2	1	0	0	0	0	0	0
1	26	40	5	2	1	3	0	0	0	0	0	0	0
2	9	17	10	4	2	1	0	0	0	0	0	0	0
3	3	6	7	10	2	3	2	1	0	0	0	0	0
4	1	7	3	6	2	3	2	1	0	0	0	1	0
5	0	3	5	4	0	4	1	1	0	0	0	0	0
6	0	2	4	1	6	1	2	0	0	0	0	0	0
7	0	2	2	0	2	0	0	0	0	0	0	0	0
8	0	0	0	2	2	0	1	2	1	0	0	0	1
9	0	0	0	1	0	0	0	1	1	0	0	0	0
10	0	0	0	0	0	1	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	2	0	0	2	0	0	0	0
13	0	0	0	0	0	1	1	0	0	0	0	0	0
14	0	0	0	0	1	0	0	0	0	0	1	0	0
15	0	0	0	0	0	0	0	0	0	0	0	1	0

Table 1. Validation data on miscounting. Shown is the total number of centrally adjudicated exacerbations per patient (columns) by the total number of self-reported exacerbations per patient (rows).

Appendix A

A negative binomially distributed random variable $y \sim \mathsf{NB}(\mu, \theta)$ can be parameterized in various ways. Here, we used

$$\mathsf{Pr}(y = k) = \frac{\Gamma(\theta + k)}{k!\,\Gamma(\theta)} \left(\frac{\theta}{\theta + \mu}\right)^{\theta} \left(\frac{\mu}{\theta + \mu}\right)^{k} , \quad k = 0, 1, 2, \dots .$$

The parameter θ thus accounts for overdispersion, with smaller θ leading to more overdispersion, and the Poisson distribution is the limiting distribution for $\theta \to \infty$. In the presence of an inflated number of zeroes, the NB distribution can be generalized to a ZINB distribution, given as

$$\mathsf{Pr}(y = k) = \begin{cases} p + (1-p)\left(\frac{\theta}{\theta+\mu}\right)^{\theta} , & k = 0 \\ (1-p)\frac{\Gamma(\theta+k)}{k!\,\Gamma(\theta)}\left(\frac{\theta}{\theta+\mu}\right)^{\theta}\left(\frac{\mu}{\theta+\mu}\right)^{k} , & k = 1, 2, \dots \end{cases}$$

with mean $\mathsf{E}(y) = (1-p)\mu$ and variance $\mathsf{Var}(y) = (1-p)\mu(1 + p\mu + \mu/\theta)$. Note that for $p = 0$ the NB distribution is retrieved, and if in addition $\theta \to \infty$, the distribution reduces to Poisson.

Appendix B

Consider the hierarchical model (6)-(7), for simplicity (and without loss of generality) with $\boldsymbol{\beta}_z = \mathbf{0}$, zero offset ($o_i = 0$ for all i), and id link for the error model. Let us start with the case without zero-inflation, *i. e.* $p_i = 0$. The rate ratio of the true treatment effect is then given as

$$\frac{\exp(\beta_0 + \beta_x)}{\exp(\beta_0)} = \exp(\beta_x) ,$$

whereas the expected value of the naive estimate is

$$\frac{\gamma_0 + \gamma_1 \cdot \exp(\beta_0 + \beta_x)}{\gamma_0 + \gamma_1 \cdot \exp(\beta_0)} \ .$$

To show that the naive estimate is always attenuated, *i. e.* biased towards 1, we distinguish two cases.

Case 1: $\beta_x < 0$

Then $\exp(\beta_x) < 1$, and therefore

$$\gamma_0 > \gamma_0 \exp(\beta_x) \ .$$

Adding $\gamma_1 \exp(\beta_0 + \beta_x)$ on both sides gives

$$\begin{aligned} \gamma_0 + \gamma_1 \exp(\beta_0 + \beta_x) &> \gamma_0 \exp(\beta_x) + \gamma_1 \exp(\beta_0 + \beta_x) \\ &= \exp(\beta_x) \cdot (\gamma_0 + \gamma_1 \exp(\beta_0)) \ , \end{aligned}$$

so

$$\frac{\gamma_0 + \gamma_1 \cdot \exp(\beta_0 + \beta_x)}{\gamma_0 + \gamma_1 \cdot \exp(\beta_0)} > \exp(\beta_x) \ , \tag{20}$$

which shows that the naive estimate of β_x is biased upwards. Moreover, the naive estimate is bounded by 1, which shows that it is biased towards 1, thus β_x is biased towards 0.

Case 2: $\beta_x > 0$

In this case, $\exp(\beta_x) > 1$, and exactly inverted arguments as in case 1 show that the naive estimate of β_x now lies between 1 and $\exp(\beta_x)$, *i. e.* is biased towards

1. We thus have attenuation of the rate ratio.

All these considerations need to be extended to the case when zero-inflation is present. The expected values $\mathsf{E}(y_i^\star \mid y_i)$ of the naive estimates must then be multiplied by $(1-p_i)$, where p_i is the zero-inflation probability for individual i. If the zero-inflation probabilities p_i were independent on the true counts y_i, so that $p_i = p$ for all i, the expected values of the naive model need to be multiplied by $(1-p)$ in the nominator and the denominator of (20), so that inequality is then still correct. The effect of non-differential error in the case with constant p is thus still attenuation.

Finally, let us look at the case with zero-inflation probabilities p_i that follow model $\mathsf{logit}(p_i) = \delta_0 + \delta_1 \mathsf{I}(y_i > 0)$ as in the TORCH study. We take the realistic assumption that the occurrence for excessive zeroes decreases (or at least does not increase) for larger true counts y_i, *i. e.* that $\delta_1 \leq 0$, which is fulfilled here, see the respective prior means in (8), (13) and (15). Going back to case 1 above, assuming $\beta_x < 0$, we have that the expected number of counts is smaller for a patient without treatment, so that also the probability that excessive counts are reported decreases for treated patients, *i. e.* $p_i^{(1)} \leq p_i^{(0)}$, with $p_i^{(1)}$ indicating the zero-inflation probability for a patient with treatment and $p_i^{(0)}$ the respective value for a patient without treatment. Therefore $1 - p_i^{(1)} \geq 1 - p_i^{(0)}$, and thus

$$\frac{(1-p_i^{(1)})(\gamma_0 + \gamma_1 \cdot \exp(\beta_0 + \beta_x))}{(1-p_i^{(0)})(\gamma_0 + \gamma_1 \cdot \exp(\beta_0))} \geq \frac{\gamma_0 + \gamma_1 \cdot \exp(\beta_0 + \beta_x)}{\gamma_0 + \gamma_1 \cdot \exp(\beta_0)} > \exp(\beta_x) \ ,$$

where the last inequality was taken from (20).

Again, for case 2 with $\beta_x > 0$ simply invert the arguments.

The above assumption $\delta_1 \leq 0$, although reasonable in our application to the TORCH study, is critical to show that non-differential miscounting error induces an attenuation effect in the hierarchical error model used here. In fact, if observed counts were generated artificially according to a ZINB model with zero-inflation $\delta_1 > 0$ (*i. e.* more zero-inflation for larger true counts), one can construct cases with overestimated treatment effects, thus reverse attenuation.